Multimodal Fusion for Depression Prediction: Exploring Intra- and Inter-Modality Dynamics

Anonymous ACL submission

Abstract

Depression prediction using clinical interviews is problematic due to small sample numbers, class imbalance, and missing modalities. The English E-DAIC corpus illustrates these limitations by offering multimodal recordings accompanied by PHQ-8 scores. We propose a hierarchical fusion approach for regression that initially enhances individual modalities by amalgamating handmade descriptors (e.g., eGeMAPS, OpenFace cues) with deep embeddings (e.g., BERT, VGGish), followed by their integration via attention-based inter-modal fusion.

Our research presents three key contributions. We present the inaugural systematic application of intra- and inter-modal fusion for regression in English clinical interviews, building upon previous research that focused on categorization or non-English datasets. Secondly, we perform a definitive robustness assessment in the presence of absent modalities, reinterpreting bimodal and trimodal results to measure modality significance and durability when data streams are deficient. Third, we illustrate that hierarchical fusion enhances generalizability in small, imbalanced clinical datasets, consistently surpassing robust baselines across MAE, RMSE, R^2, and CCC. Collectively, these results confirm structured multimodal regression as a dependable method for low-resource clinical environments and provide a foundation for interpretable, robust mental health artificial intelligence.

Keywords: Multimodal Fusion, Depression Prediction, Regression, Low-Resource Clinical Data, Intra-Inter Modality

1 Introduction

Over 280 million people are affected by depression, which is one of the main causes of disability globally. Timely prediction and evaluation of depression are essential for directing treatment and enhancing patient outcomes. Traditional diagnostic methods, like clinical interviews and PHQ-8 surveys, depend significantly on expert assessment, which may be subjective and resource-demanding. Automated methods utilizing multimodal data from clinical interviews—textual information, auditory prosody, and visual behaviors—present an opportunity to standardize depression assessment in a consistent and accessible way, especially in resource-limited settings.

Recent studies on multimodal learning have demonstrated significant potential for classification (healthy versus depressed) and regression (severity prediction) problems. Binary detection can identify danger, whereas regression facilitates a more nuanced assessment of symptom intensity, so aligning more effectively with clinical decision-making. Multimodal regression presents significant challenges in low-resource corpora like the English Distress Analysis Interview Corpus (E-DAIC), which is limited in size, demonstrates class imbalance, and has absent modalities for specific individuals. These settings require frameworks that are both precise and resilient to sparse and diverse data.

We propose the Intra- and Inter-Modal Fusion for Depression Prediction(IIFDP) framework for regression-based utilizing the E-DAIC dataset. At the intra-modal level, we integrate pre-defined descriptors (e.g., OpenSMILE audio features, OpenFace facial cues) with deep embeddings (e.g., RoBERTa/BERT for text, VGGish for audio, OpenFace for video) to include both low-level behavioural indicators and high-level semantic representations. At the inter-modal level, we utilize attention-based fusion to concurrently simulate interactions among text, audio, and visual modalities. We conduct a comprehensive evaluation of the framework across unimodal, bimodal, and multimodal situations, benchmarking it against baseline and state-of-the-art models,

utilizing regression measures such as MAE, RMSE, R^2, and Concordance Correlation Coefficient (CCC).

The primary contributions of this study are as follows:

- Initial systematic integration of intra- and inter-modal fusion for regression in English clinical interviews.
- Thorough assessment of robustness in the absence of modalities inside resource-constrained environments.
- Hierarchical fusion enhances generalizability in small, imbalanced clinical datasets.

By emphasizing regression instead of solely binary classification, our research offers a more clinically significant methodology for multimodal depression evaluation, with implications for socially relevant contexts.

2 Related Work

To contextualize our contribution, it is imperative to examine previous research on automatic depression prediction and multimodal learning. Current research encompasses a broad range, from unimodal models that depend on verbal, acoustic, or visual indicators to multimodal frameworks that amalgamate many behavioral inputs. Recent methodologies have started to prioritize hierarchical fusion techniques, integrating handcrafted descriptors with deep embeddings both within and across modalities. The limited resources of clinical corpora like E-DAIC highlight the necessity for techniques that are both resilient and socially significant. This study presents a summary of pertinent research in unimodal, multimodal, and fusion-based frameworks, while identifying deficiencies that justify our regression-oriented adaption of intra- and inter-modal fusion.

2.1 Unimodal Approaches for Depression Detection

Initial studies on automatic depression diagnosis investigated unimodal data sources, typically concentrating on a singular behavioural channel. Text-based methodologies utilize language indicators, including sentiment, lexical diversity, and psycholinguistic patterns, to forecast the severity of depression. The advent of deep learning has enhanced regression performance through contextual embeddings from models such as BERT and RoBERTa, which effectively capture nuanced language usage. Acoustic techniques utilize manually built features such as eGeMAPS or MFCCs in conjunction with deep encoders like VGGish, as speech prosody and vocal quality are significant indicators of emotional states. Visual methodologies extract non-verbal behaviours, including facial action units, eye gaze, and head attitude, utilizing tools such as OpenFace, while temporal models (CNNs, LSTMs, or Transformers) are employed to record temporal progression. Although successful independently, unimodal models frequently exhibit low resilience and generalization due to their inadequate depiction of depressed symptoms, which are intrinsically multimodal.

2.2 Multimodal Integration in Depression Detection

To overcome the constraints of unimodal approaches, current research has introduced multimodal frameworks that integrate textual, audio, and visual modalities. Early fusion methods concatenate feature vectors from different modalities, whereas late fusion integrates independent predictions. Nonetheless, mere concatenation frequently results in redundancy and diminished performance. Advanced fusion solutions encompass attention-based architectures, recurrent neural networks, and Transformer-based designs. For instance, MulT presented a cross-modal Transformer to describe long-range relationships across modalities, whereas CubeMLP redefined fusion as learning along the sequence, modality, and channel dimensions. PMR (Progressive Modality Reinforcement) has implemented a message hub to enhance underrepresented modalities, utilizing self-attention or cross-attention techniques to highlight relevant cues. These studies illustrate the advantages of recording cross-modal interactions, but the majority emphasize classification over regression problems.

2.3 Intra- and Inter-Modal Fusion Frameworks

A significant difficulty, in addition to cross-modal fusion, is the integration of complementary elements within a single modality. Handcrafted descriptors (Jin et al., 2024) (e.g., OpenSMILE (Eyben et al., 2010), OpenFace) offer interpretable low-level indicators (Zhoua and Ganb, 2008), whereas

deep embeddings (Ustinova and Lempitsky, 2016) (e.g., BERT, VGGish, Pose, Gaze, AU) encapsulate high-level semantics. Recent research, including the IIFDD framework (Chen et al., 2024), introduced intra-modal fusion transformers to integrate handmade and deep representations prior to executing inter-modal fusion. This hierarchical structure guarantees that the representation of each modality is enhanced prior to cross-modal integration, resulting in improved performance. Although promising, previous studies like IIFDD were predominantly conducted on Chinese corpora (CMDC, EATD), raising the question of the efficacy of intra-/inter-modal fusion on English clinical datasets such as E-DAIC. Our research fills this void by modifying IIFDD for regression analysis on English interviews.

2.4 Perspectives on Low Resources and Social Impact

A persistent difficulty in depression diagnosis is the resource scarcity of datasets: restricted sample sizes, uneven distributions, and inadequate modalities. The E-DAIC corpus illustrates these problems, rendering it an optimal instance for resilient multimodal frameworks. To alleviate shortages, previous research has investigated data augmentation, transfer learning, and modality dropout techniques. From a comprehensive viewpoint, multimodal depression detection also relates to social implications: developing equitable, transparent, and clinically dependable models that can assist healthcare practitioners and enhance access for marginalized people. The integration of interpretable fusion processes promotes performance while also fostering trustworthiness and ethical deployment.

3 Methodology

Our objective is to develop a methodology capable of accurately predicting depression ratings from multimodal clinical interviews in low-resource settings. To do this, we enhance the intra- and inter-modal fusion paradigm and tailor it to the English E-DAIC dataset, assuring the effective utilization of both handmade descriptors and deep embeddings inside and across modalities, where Figure 1 depicts the comprehensive architecture of our proposed intra- and inter-modal fusion framework for regression utilizing the E-DAIC dataset.

3.1 Dataset

Our research is founded on the English Distress Analysis Interview Corpus (E-DAIC) (Gratch et al., 2014) (DeVault et al., 2014) (Ringeval et al., 2019), comprising semi-structured clinical interviews conducted between participants and a virtual agent. Every participant receives a PHQ-8 score, facilitating both binary depression identification and regression-based severity forecasting. The dataset is fundamentally low-resource: participant numbers are restricted, the distribution of PHQ-8 scores is uneven, and video recordings are absent for a subset of users. The attributes of E-DAIC provide it a suitable platform for assessing resilient multimodal fusion methodologies.

3.2 Feature Extraction

To obtain both superficial and profound representations of participant behaviour, we extract characteristics from three modalities:

- **Textual Features:** Transcribed transcripts are integrated via BERT and NLP based sentiment to encapsulate contextual semantics. Moreover, linguistic metrics like sentence length and word-level statistics are deemed to preserve manual interpretability.
- **Acoustic Features:** We extract eGeMAPS descriptors from auditory responses with OpenSMILE, which reflect prosodic and spectral elements. Concurrently, VGGish embeddings, pre-trained on extensive audio datasets, are utilized to capture advanced acoustic semantics.
- **Visual Features:** Raw video frames are absent, and features are analyzed using OpenFace to extract facial action units, gaze direction, and head orientation separately as manually produced indicators.

This dual extraction technique offers two complementary perspectives for each modality: interpretable handcrafted descriptors and semantically enriched embeddings.

3.3 Intra-Modal Fusion

A primary problem is that handmade and deep representations encapsulate distinct yet complementary aspects of the same modality. Addressing them in isolation disregards potentially significant insights. To resolve this, we employ intra-modal

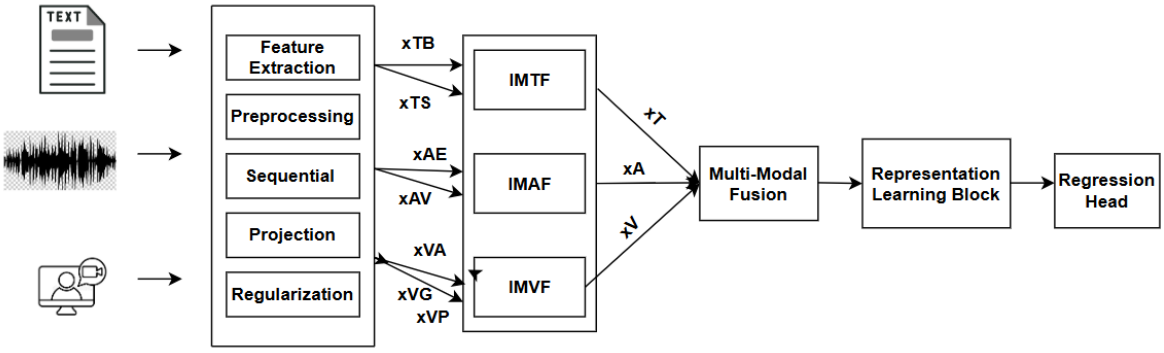

Figure 1: Overview of the proposed hierarchical intra- and inter-modal fusion framework. Text, audio, and visual features are first enriched via intra-modal fusion (IMTF, IMAF, IMVF), followed by multi-modal fusion, representation learning, and regression prediction.

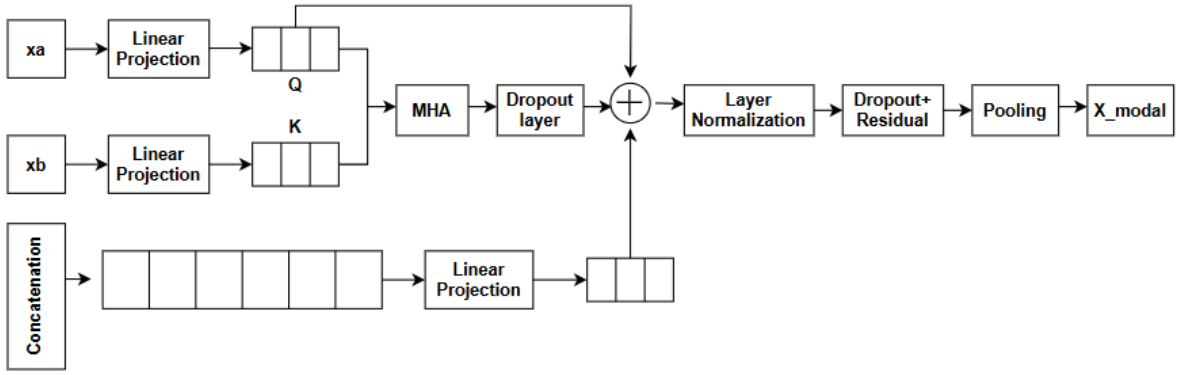

Figure 2: Architecture of the proposed cross-modal attention module. Inputs (x_a, x_b) are projected, transformed via multi-head attention (MHA), normalized, and aggregated with residual connections. The pooled output X_{modal} forms the fused representation.

fusion transformers. Initially, the handcrafted features and deep embeddings for each modality are projected into a shared latent space. Attention mechanisms are employed to synthesize them, guaranteeing that low-level signals enhance high-level embeddings and vice versa. Consequently, each modality produces a singular enriched representation that harmonizes interpretability with expressive capability.

3.4 Inter-modal Fusion

Intra-modal fusion enhances individual modalities, whereas depression manifests through interactions among language, voice, and facial expressions. We utilize inter-modal fusion through attention techniques to capture these dependencies. Pairwise bimodal fusion modules initially align and integrate modality-specific encodings (e.g., text-audio, audio-video, text-video). These are then consolidated in a tri-modal attention module that learns weighted combinations of the three modalities. This hierarchical design, in contrast to basic concatenation, enables the model to prioritize modality interactions that are most indicative of depression prediction, while diminishing the influence of redundant input.

4 Experimental Setup

We performed extensive tests on the English E-DAIC dataset to verify the efficacy of the suggested framework. This part talks about the dataset splits, preparation steps, training settings, baseline models, and assessment measures. By explaining these parts in detail, we make sure that our results can be repeated and give a fair way to compare them with other methods for predicting multimodal depression.

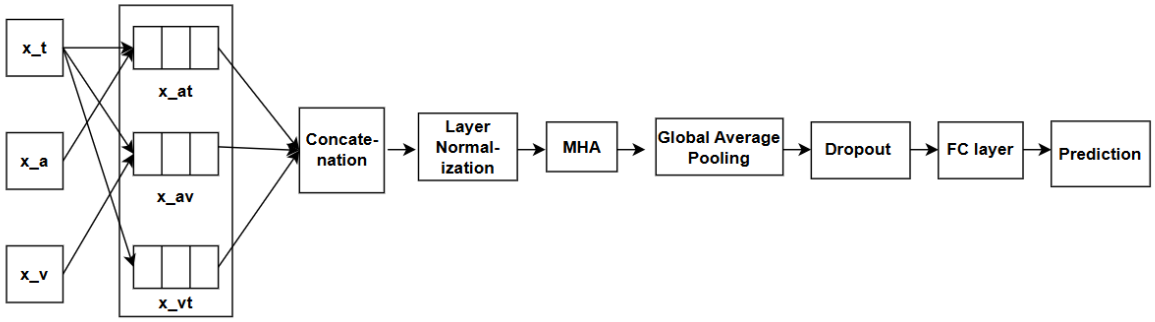

Figure 3: Overview of the proposed multimodal attention-based framework. Text (x_t), audio (x_a), and visual (x_v) features interact pairwise (x_{at}, x_{av}, x_{vt}), followed by concatenation, normalization, multi-head attention (MHA), global pooling, dropout, and a fully connected layer for prediction.

4.1 Data Splits

We adhere to the approved E-DAIC (Gratch et al., 2014) (Ringeval et al., 2019) (DeVault et al., 2014) protocol, which are a total of 275 users, categorizing individuals into three groups: 163 for training, 56 for development, and 56 for testing. Each division maintains the subjects' separation, ensuring that no participant appears in multiple batches. The training set identifies optimal values for the model's parameters, the development set fine-tunes hyperparameters and facilitates early stopping, while the test set is reserved solely for final evaluation. This article utilizes just the approved test set to ensure comparability with prior results.

4.2 Data Preprocessing

Prior to modelling, each modality is subjected to preprocessing to guarantee uniformity, where text-transcribed transcripts are sanitized, tokenized, and embedded with pre-trained BERT models. Sentence-level embeddings are aggregated to derive participant-level vectors. Audio recordings are divided and analyzed with OpenSMILE to extract eGeMAPS features, while deep embeddings are produced using VGGish. Features are standardized by z-score normalization. Visual attributes are derived from frame sequences utilizing OpenFace, generating action units, gaze, and head attitude indicators.
All features are organized by participant_ ID to establish a uniform representation across modalities.

4.3 Baseline Standards

To evaluate the efficacy of our system, we compare it against multiple recognized baselines: GRU-BiLSTM (Shen et al., 2022) and BiLSTM (Zou et al., 2022) are robust sequence models for temporal features. MulT (Zou et al., 2022)is a Transformer-based cross-modal model(Rajan et al., 2022). CubeMLP (Sun et al., 2022), and PMR (Lv et al., 2021) are the architecture for modality-sequence-channel fusion.
These baselines were selected as they exemplify distinct models of multimodal learning: recurrent, transformer-based, and MLP-driven fusion.

4.4 Training Procedure

All models encompassing the proposed framework and baseline architectures were trained under uniform conditions to provide an equitable comparison. Training utilized the the Adam optimizer. with the weight decay and learning rate, n was adjusted within the interval of $[1 * 10^{-}5, 5 * 10^{-}4]$ according to the performance on the development and training set. A batch size of 16 to 32 was utilized based on modality dimensionality, and the training duration was limited to 200 epochs. To alleviate overfitting, we implemented early stopping with a patience of 10 epochs, alongside dropout layers and batch normalization inside the model architecture. The optimization aim was the Mean Squared Error (MSE) loss, defined as

$$\text{LMSE} = \frac{1}{B}\sum_{i=1}^{B}(y_i - \hat{y}_i)^2$$

where $\hat{y}$

signifies the actual PHQ-8 score and the anticipated score for each sample in a batch of size B. All tests were conducted on NVIDIA GPUs, facilitating rapid training of multimodal models while ensuring computational consistency across trials.

4.5 Evaluation Metrics

To thoroughly assess the regression efficacy of our framework in predicting PHQ-8 severity, we utilize four commonly employed metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Concordance Correlation Coefficient (CCC). Collectively, these criteria evaluate the precision, resilience, and dependability of the forecasts.

5 Results and Discussion

The proposed framework was assessed in unimodal, bimodal, and multimodal situations utilizing the criteria delineated in Section 4.5. This path enables us to examine the evolution of performance as features are incorporated inside and across modalities and to assess our methodology against established baselines.

5.1 Unimodal and Intra-Modal Fusion Results

The unimodal results in Table 1 illustrate the baseline effectiveness of individual modalities and the influence of intra-modal feature fusion. Of the three modalities, text exhibits superior performance in regression measures, with RoBERTa-based embeddings (BERT and XM) utilized alongside recurrent models like LSTM and GRU-BiLSTM attaining the lowest MAE and RMSE values, while regularly sustaining CCC scores exceeding 0.8. Attention-based baselines (Vaswani et al., 2017), encompassing self- and cross-attention, significantly augment agreement scores, highlighting the importance of contextual weighting in the enhancement of textual embeddings. The auditory modality demonstrates increased variability, with VGG-based deep embeddings outperforming handcrafted eGeMAPS features, suggesting that pretrained deep acoustic representations more adeptly capture clinically relevant prosodic patterns. Significantly, the intra-modal fusion of VGG and eGeMAPS diminishes error and enhances CCC relative to each feature type in isolation, indicating that handmade descriptors, although less effective independently, offer additional insights when combined with deep embeddings. In contrast, the visual modality exhibits constrained predictive capability, as characteristics related to stance, gaze, and action units produce comparatively elevated MAE and RMSE values, with CCC scores often falling below 0.5. Nonetheless, combining these visual characteristics produces measurable improvements over isolated streams, indicating that while no single visual channel is strongly predictive, their joint modeling yields a more coherent signal of behavioral patterns.

5.2 Bimodal and Trimodal Fusion Results

The outcomes of bimodal and multimodal fusion, as presented in Table 2, demonstrate consistent enhancements compared to unimodal baselines. Among bimodal pairings, text–audio has the highest performance, decreasing prediction error and elevating CCC above 0.75, indicating that prosodic variations enhance language semantics in identifying sad speech patterns. Text-visual fusion yields enhancements, but less significantly, and audio-visual fusion is advantageous yet stays inferior to text-driven combinations.

The integration of all three modalities through the proposed intra- and inter-modal fusion design (IIMD) yields superior overall results, surpassing baseline models including GRU-BiLSTM, BiLSTM, MulT, and CubeMLP. The trimodal configuration produces the lowest error rates and the highest agreement metrics, especially in CCC, indicating enhanced concordance with actual severity levels. These findings validate that hierarchical fusion not only utilizes the advantages of text but also incorporates supplementary audio and visual elements, providing a stable and dependable framework for PHQ-8 regression in resource-limited clinical settings.

6 Conclusion and Future Work

This study presented a hierarchical architecture for intra- and inter-modal fusion aimed at regression-based depression prediction utilizing the English E-DAIC dataset. By methodically integrating handcrafted descriptors with deep embeddings at the intra-modal level and modeling cross-channel interdependence through inter-modal attention, our methodology successfully surpassed robust baselines across MAE, RMSE, R^2, and CCC metrics.

Our investigation yielded three principal insights. Initially, intra- and inter-modal fusion can be proficiently applied to English clinical interviews for regression, so filling a void in previous research. The study of robustness in the absence of certain modalities indicates that text is the primary modality, although auditory and visual signals offer supplementary resilience. Third, hierarchical fusion enhances generalizability in tiny, imbalanced datasets, pro-

Table 1: Unimodal and intra-modal fusion results on E-DAIC. Text is the strongest modality; intra-modal fusion improves audio and visual streams.

Modal	Features	Metrics	GRU-BiLSTM	BiLSTM	LSTM	MulT	PMR	Self-att	Cross-att	CubeMLP	Conv1D
T	bert	MAE	2.56	2.78	**1.49**	-	-	2.21	-	-	2.67
		RMSE	3.15	2.95	**2.14**	-	-	2.75	-	-	3.31
		R^2	0.64	0.70	**0.87**	-	-	0.81	-	-	0.65
		CCC	0.82	0.78	**0.93**	-	-	0.90	-	-	0.87
	xm	MAE	5.25	**2.67**	2.83	-	-	4.544	-	-	3.82
		RMSE	5.41	3.78	**3.73**	-	-	5.80	-	-	4.87
		R^2	**0.67**	0.66	0.66	-	-	0.281	-	-	0.41
		CCC	0.56	**0.83**	0.78	-	-	0.578	-	-	0.77
	bert+xm	MAE	**2.88**	3.11	2.86	4.12	5.19	3.02	3.29	4.89	3.05
		RMSE	**3.55**	3.79	3.98	6.89	6.48	3.84	4.09	5.64	3.80
		R^2	0.68	**0.79**	0.70	0.64	0.59	0.63	0.59	0.73	0.64
		CCC	0.80	0.84	0.81	0.75	0.62	0.77	0.45	**0.86**	0.81
A	Vggish	MAE	**2.77**	2.96	2.98	-	-	5.27	-	-	5.34
		RMSE	3.36	**3.32**	3.86	-	-	6.54	-	-	6.54
		R^2	0.72	**0.75**	0.63	-	-	0.15	-	-	0.08
		CCC	0.84	0.88	0.78	-	-	0.28	-	-	0.21
	EgeMAPS	MAE	3.74	3.91	4.98	-	-	4.60	-	-	5.67
		RMSE	4.71	5.00	6.02	-	-	5.80	-	-	6.77
		R^2	0.459	0.38	0.28	-		0.17	-	-	-0.13
		CCC	0.64	0.56	0.21	-	-	0.33	-	-	0.01
	Vggish+ EgeMAPS	MAE	5.18	4.95	5.21	5.34	6.99	7.59	5.38	5.26	4.14
		RMSE	6.24	6.03	6.34	7.48	8.24	8.47	6.66	6.45	5.01
		R^2	0.15	0.10	0.34	-0.03	0.25	0.24	0.48	-0.10	0.55
		CCC	0.29	0.14	0.56	0.11	0.21	0.35	0.66	0.22	0.62
V	Pose		5.63	6.27	4.18	-	-	2.73	-	-	11.93
		RMSE	6.88	8.01	5.25	-	-	4.14	-	-	23.65
		R^2	0.42	0.38	0.54	-	-	0.57	-	-	-12.76
		CCC	0.35	0.24	0.48	-	-	0.70	-	-	0.04
	Gaze	MAE	5.41	2.34	3.01	-	-	2.09	-	-	3.12
		RMSE	6.02	3.00	3.92	-	-	3.40	-	-	4.10
		R^2	0.54	0.65	0.62	-	-	0.70	-	-	0.58
		CCC	0.69	0.88	0.73	-	-	0.83	-	-	0.74
	AU	MAE	3.58	3.31	4.41	-	-	3.33	-	-	5.75
		RMSE	4.84	4.40	4.72	-	-	4.51	-	-	7.12
		R^2	0.42	0.52	0.45	-	-	0.62	-	-	-0.24
		CCC	0.67	0.71	0.63	-	-	0.85	-	-	0.04
	Pose+ Gaze+AU	MAE	5.34	4.82	4.12	4.88	6.29	5.56	3.98	6.98	6.12
		RMSE	6.50	6.92	5.59	5.73	8.48	6.67	4.56	8.52	6.10
		R^2	-0.04	-0.05	0.32	0.27	-0.25	0.41	0.38	-0.12	0.20
		CCC	0.19	0.29	0.37	0.15	0.05	0.45	0.55	0.10	0.34

Table 2: Bimodal and trimodal results reframed as robustness analysis. Removing text causes the largest drop, while audio and visual provide complementary resilience.

Modal	Feature	Metric	GRU-BI LSTM	BI LSTM	LSTM	MulT	PMR	Self-att	Cross-att	Cube MLP	IIMD
A+T	Fusion	MAE	5.18	5.22	4.89	4.74	4.66	5.27	6.01	5.23	**4.18**
		RMSE	6.24	6.28	5.87	5.68	5.62	6.44	8.62	6.18	**5.25**
		R^2	0.05	0.27	0.15	0.49	0.14	-1.62	0.41	0.41	**0.51**
		CCC	0.12	0.55	0.23	0.54	0.27	0.34	0.48	0.56	**0.59**
A+V	Fusion	MAE	3.85	4.71	2.76	5.43	5.31	6.45	5.70	5.46	**2.91**
		RMSE	4.79	4.25	4.35	6.49	7.45	8.89	6.97	6.53	**3.84**
		R^2	0.41	0.73	0.57	-0.03	-0.01	-0.11	-0.03	0.35	**0.62**
		CCC	0.32	0.74	0.26	0.11	0.19	0.21	0.10	0.15	**0.79**
V+T	Fusion	MAE	4.46	3.53	5.38	4.69	4.32	4.88	4.77	5.46	**3.48**
		RMSE	5.59	**4.47**	6.55	5.81	5.50	5.75	5.74	6.53	4.79
		R^2	0.23	0.50	**0.57**	0.16	0.19	0.38	0.18	0.35	0.28
		CCC	0.39	**0.67**	0.19	0.27	0.32	0.21	0.29	0.15	0.46
A+V+T	Fusion	MAE	3.91	4.73	5.34	4.50	5.32	3.59	6.06	6.89	3.53
		RMSE	6.00	5.61	6.47	5.21	6.55	4.57	8.58	8.41	4.79
		R^2	0.11	0.17	0.21	0.52	0.05	0.22	0.49	0.29	0.38
		CCC	0.15	0.54	0.14	0.58	0.02	0.10	0.15	0.19	0.40

viding a more dependable alternative to simplistic concatenation methods.

Future endeavors will thus focus on three avenues: (i) implementing modality-dropout during training to explicitly bolster robustness against incomplete inputs, (ii) broadening validation across datasets like AVEC to ensure greater generalizability, and (iii) incorporating explainability mechanisms—such as attention visualization, feature attribution, and clinician-informed thresholds—to improve trust and usability in clinical practice.

Collectively, these guidelines will facilitate the creation of multimodal depression prediction systems that are accurate, reliable, interpretable, and congruent with therapeutic requirements.

References

Jian Chen, Yuzhu Hu, Qifeng Lai, Wei Wang, Junxin Chen, Han Liu, Gautam Srivastava, Ali Kashif Bashir, and Xiping Hu. 2024. Iifdd: Intra and inter-modal fusion for depression detection with multi-modal information from internet of medical things. *Information Fusion*, 102:102017.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, and 1 others. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pages 3123–3128. Reykjavik.

Chenyu Jin, Shuchang Zhao, Shiqing Zhang, Zhewei Fang, Junjie Xie, and Ying Chen. 2024. Attention-based audio depression recognition integrating handcrafted and deep features. In *CSIG Conference on Emotional Intelligence*, pages 206–218. Springer.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2554–2562.

Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. 2022. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4693–4697. IEEE.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, and 1 others. 2019. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3722–3729.

Evgeniya Ustinova and Victor Lempitsky. 2016. Learning deep embeddings with histogram loss. *Advances in neural information processing systems*, 29.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shang-Ming Zhoua and John Q Ganb. 2008. Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modeling. *Fuzzy Sets and Systems*, 159:3091–3131.

Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. 2022. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4):2823–2838.